Amendments to the Claims

1. (currently amended) A computer-implemented method for information retrieval, classification, indexing, and summarization, comprising:

identifying a collection of hyperlinked documents as a single coherent compound document on a single topic created by a number of collaborating authors, wherein the identifying includes observing results of a first number of heuristics run on the collection of hyperlinked documents and related hyperlinks, ~~wherein the first number of heuristics includes identifying at least one of: similar creation dates and similar last-modified dates;~~, and wherein the collection of hyperlinked documents is distributed over a plurality of URLs, wherein the first number of heuristics includes:

identifying hyperlinks that link within a same directory and include a sufficient quantity of common anchor text,

identifying hyperlinks that contain linguistic structures that indicate relationships between document parts,

identifying external hyperlinks to same places,

identifying at least one of: similar creation dates and similar last-modified dates,

identifying individual URLs having similar structure indicating an order of inclusion in the compound document, and

identifying a link structure of "wheel" form;

analyzing ~~the~~ content and structure of the compound document to find a preferred entry point for the compound document, wherein the analyzing includes observing results of a second number of heuristics run on the compound document and related hyperlinks~~, wherein the analyzing includes~~ and combining the results of the second number of heuristics run on various hyperlinked documents of the compound document, wherein the results of the second number of heuristics include numerical scores, wherein ~~and~~ the combining includes a weighted averaging of the numerical scores into an overall score, ~~and~~ wherein a maximum overall score determines the preferred entry point~~;~~, and wherein the second number of heuristics includes:

identifying specific filenames that define entry points, including at least one of: "index" and "default",

identifying a particular component document in the compound document as a suitable entry point because the component document has several in-links, wherein the in-links are from outside the compound document,

determining a measure of vector distances along intra-document links between a particular component document and all other component documents in the compound document,

determining whether a URL has links pointing to longer URLs having common directory components followed by different ending directory components, wherein the ending directory components contain specific identifying information;

processing the compound document as a whole, wherein processing the compound document as a whole includes ~~including~~ at least one of: indexing, classification, and retrieval; and

processing the compound document from the entry point, wherein processing the compound document from the entry point includes ~~including at least one of~~ creating ~~at least one of~~ a presentation of ~~results from retrieval, summarization, and classification.~~ the entry point as a URL.

2. (currently amended) The method of claim 1 wherein the collection of hyperlinked documents includes material from at least one of: ~~the~~ an internet, an intranet, and a digital library.

3-5. (cancelled)

6. (currently amended) The method of claim 1 wherein the first number of heuristics includes identifying hyperlinks that contain linguistic structures that indicate relationships between parts of a document including at least one of: a list of page

numbers, ~~the~~ terms "next", "previous", "index", and "contents", and the terms' ~~their~~ non-English equivalents.

7-14. (canceled)

15. (previously presented) The method of claim 1 wherein the second number of heuristics includes identifying a particular component document in the compound document as the entry point because the component document has several out-links.

16-21. (canceled)

22. (new) The method of claim 1, further comprising:

prior to identifying the collection of hyperlinked documents as the single coherent compound document on the single topic created by the number of collaborating authors, processing a data set of URLs representing the collection of hyperlinked documents to remove URLs which do not qualify as compound documents.

23. (new) The method of claim 22 wherein processing the data set of URLs includes filtering out URLs from the data set with an HTTP return code of 400 or greater.

24. (new) The method of claim 22 wherein processing the data set of URLs includes removing URLs from the data set with a URL extension that does not contain textual content.

25. (new) The method of claim 22 wherein processing the data set of URLs includes resolving redirects within a directory.

26. (new) The method of claim 22 wherein processing the data set of URLs includes filtering out URLs from the data set with an HTTP return code of 400 or greater,

removing URLs from the data set with a URL extension that does not contain textual content, and resolving redirects within a directory.

27. (new) The method of claim 1, further comprising:

prior to identifying the collection of hyperlinked documents as the single coherent compound document on the single topic created by the number of collaborating authors, processing a data set of URLs representing the collection of hyperlinked documents, wherein processing the data set of URLs includes removing an argument part of a URL.

28. (new) The method of claim 27 wherein the argument part of the URL follows a # symbol or a ? symbol.

29. (new) The method of claim 27 wherein processing the data set of URLs further includes resolving redirects within a directory.

30. (new) The method of claim 27 wherein processing the data set of URLs further includes filtering out URLs from the data set with an HTTP return code of 400 or greater.

31. (new) The method of claim 27 wherein processing the data set of URLs further includes removing URLs from the data set with a URL extension that does not contain textual content.

32. (new) The method of claim 1, further comprising:

prior to identifying the collection of hyperlinked documents as the single coherent compound document on the single topic created by the number of collaborating authors, processing a data set of URLs representing the collection of hyperlinked documents, wherein processing the data set of URLs includes resolving redirects within a directory, filtering out URLs from the data set with an HTTP return code of 400 or greater, removing URLs from the data set with a URL extension that does not contain textual content, and removing an argument part of a URL, wherein the argument part of the URL follows a # symbol or a ? symbol.

33. (new) The method of claim 1, further comprising:

prior to identifying the collection of hyperlinked documents as the single coherent compound document on the single topic created by the number of collaborating authors, processing a data set of URLs representing the collection of hyperlinked documents, wherein processing the data set of URLs includes ignoring a directory that contains less than 3 URLs and ignoring a directory that contains more than 250 URLs.

34. (new) The method of claim 1, further comprising:

prior to identifying the collection of hyperlinked documents as the single coherent compound document on the single topic created by the number of collaborating authors, processing a data set of URLs representing the collection of hyperlinked documents, wherein processing the data set of URLs includes ignoring a directory that contains URLs generated by an Apache.TM. web server.